DOCUMENT RESUME

ED 393 881 TM 024 761

AUTHOR Carlson, Randal D.; Locklin, Ralph H.
TITLE Item Response Theory: Comparing BILOG and MicroCAT Calibration for a Mathematics Ability Test.
PUB DATE [95]
NOTE 31p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; *College Freshmen; Comparative Analysis; Estimation (Mathematics); Higher Education; *Item Response Theory; *Mathematics Tests; *Test Items
IDENTIFIERS *BILOG Computer Program; Calibration; *MicroCAT Testing System; Missing Data; Three Parameter Model; Two Parameter Model

ABSTRACT

This study reports the results of an investigation into the accuracy and efficacy of item calibration schemes used by two commercially available personal computer programs, BILOG and MicroCAT, when used to calibrate a test that is currently used in higher education. A calibration of 1,000 randomly selected students' responses to a 72-question math examination taken by all freshmen entering a large Eastern research university was performed using various available options of the two programs. A comparison was made between the calibration schemes concerning the parameters determined, item fits, and the resulting ability estimates. High agreement was found between the programs in item parameterization. BILOG appeared to provide a better fit to the chosen parameterization model in the 2- and 3-parameter cases. Estimation of abilities was also quite similar: differences encountered were more pronounced in the estimation of the ability of low scoring examinees. The effect of using a sample with all responses complete as contrasted with a sample containing omitted responses appeared to be quite small regardless of the program used. (Contains 4 tables, 6 figures, and 21 references.) (Author/SLD)

Item Response Theory: Comparing BILOG and MicroCAT Calibration for a Mathematics Ability Test

Randal D. Carlson

Department of Educational Leadership, Technology and Research

Georgia Southern University

Ralph H. Locklin

University Testing Service

The Pennsylvania State University

Abstract

This study reports the results of an investigation into the accuracy and efficacy of item calibration schemes used by two commercially available personal computer programs, BILOG and MicroCAT, when used to calibrate a test that is currently used in higher education. A calibration of 1000 randomly selected students' responses to a 72 question math exam taken by all freshmen entering a large Eastern research university was performed using various available options of the two programs. A comparison was made between the calibration schemes concerning the parameters determined, item fits, and the resulting ability estimates. High agreement was found between the programs in item parameterization. BILOG appeared to provide a better fit to the chosen parameterization model in the 2- and 3-parameter cases. Estimation of abilities was also quite similar; differences encountered were more pronounced in the estimation of the ability of low scoring examinees. The effect of using a sample with all responses complete as contrasted with a sample containing omitted responses appeared to be quite small regardless of program used.

Statement of the Problem

For decades, Classical Test Theory (CTT) has been a useful tool that has allowed psychometricians to characterize tests and test items. Still, according to Hambleton and Swaminathan (1985), there are many documented shortcomings of this theory. Item Response Theory (IRT) addressed these problems. It has been gaining acceptance in educational testing because it provides more adaptable and effective methods of test construction, analysis, and scoring power. IRT in its simplest form treats test items as small, interchangeable units of test construction and scoring (Mislevy and Bock, 1990).

The estimation of the item parameters, discrimination (a), difficulty (b), and guessing (c), must be accomplished in order to describe each item. Although these parameters are characteristic of each item, their estimation is tied together with the ability of the test taker (Θ). These four variables are latent in that they cannot be directly measured. They must be inferred through some mathematical calculation. The only measurable characteristic is the test taker's answers to a series of questions, **u**, where **u** is a response vector the length of which equals the number of test items. Finding the parameters Θ for the test taker and a, b, and c for the item is termed the problem of "joint estimation of parameters." This problem is so named because all the parameters must be estimated simultaneously.

Three different logistic models are assumed. The most general, called the 3-parameter model involves the estimation of all 3 structural parameters (a, b, and c). The 2-parameter model fixes c at 0 and the 1-parameter or Rasch model fixes c at 0 and a at a constant value.

This study compares the solutions to the problem of joint estimation of parameters as operationalized in the commercially available programs, BILOG and MicroCAT. This comparison may help practitioners choose calibration programs for use with real data. Specifically, the following questions are addressed.

1. Are there differences in the estimated parameters computed by the two programs?
2. Do the programs produce similar parameter estimates in the presence of omitted answers?
3. Do the programs differ in the estimated ability of subjects in each sample?

4. Do the goodness-of-fit statistics of each program identify the same items for possible deletion from the test?

Theoretical Description

Calibration

MicroCAT uses two different programs to estimate item parameters: RASCAL for Rasch calibration and ASCAL for 2- and 3-parameter calibration. RASCAL is based on the Wright and Stone (1979) general unconditional calibration method. This method is conceptually simpler than other computational methods because the problem of insufficient statistics is avoided. Several modifications and enhancements were made to RASCAL to make it more consistent with the 3-parameter model. Most noticeable is the option to standardize either the item difficulties to a mean of 0 in the historical Rasch fashion or to scale the ability distribution to a mean of 0 and standard deviation of 1. If standardization of item difficulties is chosen, a correction for bias is applied (Wright and Stone, 1979). One can show by plotting the two different scaling procedures that one scale is simply a linear transformation of the other.

ASCAL uses a combined maximum likelihood and modal Bayesian procedure (Vale & Giallucia, 1985) to estimate the 2- and 3-parameter solutions. The item parameters are estimated through an iterative procedure. Initially, normal-curve approximations are used for the discrimination, difficulty, and ability parameters (Jensma, 1976) and the reciprocal of the number of response alternatives is used for the guessing parameter (if applicable). A Bayesian adaptation of Lord's (1974) maximum likelihood equations modified to allow a normal Bayesian prior for ability and a beta prior distribution for the a and c parameters is used to provide the final parameter estimates.

BILOG 3.04 (Mislevy & Bock, 1990) uses the calibration method of marginal maximum likelihood (MML). Bock and Lieberman (1970), Bock and Aitken (1981), and Hartwell, Baker, & Zwarts. (1988) suggest that the process of estimation of parameters can be improved if the ability parameter (Θ) is removed so that the process can concentrate on the structural parameters (a, b, and

c). Removal of the ability parameter, called the incidental parameter by Hambleton & Swaminathan (1985) allows the likelihood function to be expressed only in terms of the structural parameters (a, b, and c). The probability of one examinee obtaining a particular response vector, **u**, is:

$$L(\mathbf{u}|\Theta,\mathbf{a},\mathbf{b},\mathbf{c}) = \prod_{i=1}^{N} P^{u_i} Q^{(1-u_i)}.$$

If the distribution of Θ across subjects can be defined as $g(\Theta)$, it follows that integrating across all Θ allows the resultant to express the unconditional likelihood (in terms of Θ) of response pattern u_i which is commonly referred to as the as the marginal likelihood (Hambleton & Swaminathan, 1985),

$$L(\mathbf{u},\Theta|\mathbf{a},\mathbf{b},\mathbf{c}) = \int_{-\infty}^{\infty} \prod_{i=1}^{N} P^{u_i} Q^{(1-u_i)} g(\Theta) d\Theta = \pi_j.$$

In general, the integral cannot be expressed in closed form, but can be solved for as accurately as desired by using the Gausian Quadriture formula (Mislevy & Bock, 1990).

Since there are 2^N response patterns for the N binary items, the overall likelihood function can calculated. To avoid the problem of small numbers, this function is expressed as a log:

$$\ln[L(\mathbf{u}|\mathbf{a},\mathbf{b},\mathbf{c})] = k + r_j \sum_{j=1}^{2^N} \ln(\pi_j),$$

where r_j = the number of examinees obtaining response pattern u_j (Hambleton & Swaminathan, 1985) and k= a constant.

The marginal maximum likelihood estimators are obtained by differentiating this function with respect to the parameters a, b, and c. Then the resulting system of equations must be solved. The EM algorithm (Dempster, Laird, & Rubin, 1977) and Newton-Gauss (Fisher scoring) methods are used in the solution of these equations (Bock & Aitken, 1981; Thissen, 1982).

A Bayesian procedure called marginal maximum *a posteriori* (MMAP) can be used to constrain the parameters to keep them from exceeding allowable limits. This is done through the specifying of prior distributions similar to other Bayesian approaches (Mislevy, 1986; Tsutakawa & Lin, 1986).

Goodness of Fit

Goodness of fit of the model is evaluated in both programs by the calculation of a likelihood-ratio chi-square statistic. The evaluation of each item within the context of the fit of the model can be used as an aid to determine whether an item is so ill fitting that it should be discarded from the test.

RASCAL and ASCAL compute a Pearson chi-square statistic (Bock, 1975) to test for lack of fit. Expected frequencies calculated from marginal frequencies and observed frequencies are tabled for each of 20 or fewer fractiles. The usual formulas are used to compute the statistic after collapsing the fractiles to assure that there are at least five examinees per fractile.

For long tests such as the one used in this study, BILOG 3.04 (Mislevy & Bock, 1990) computes a likelihood-ratio chi-square statistic that is used to compare the frequency of correct and incorrect responses in ability intervals with those from the fitted model at the mean interval, Θ_h. The expected *a posteriori* (EAP) estimate with the same priors used for calibration is used to estimate the abilities (Θ). The Θs are then rescaled so that the variance of the sample distribution equals the latent distribution on which the MML estimation of item parameters is based. The number of subjects responding correctly in an interval are tallied and a likelihood-ratio chi-square statistic used to compare the resulting frequencies of correct and incorrect responses.

Ability Estimation

Both MicroCAT and BILOG use 3 different methods to calculate ability based on the parameters as determined in the calibration phase. They are: modal Bayesian, maximum likelihood, and expected *a posteriori* Bayesian (EAP) methods. Although not identical in the way that they are operationalized, they are quite similar in their theoretical foundations. A fuller explanation can be found in Hambleton and Swaminathan (1985).

Design

Instrument

The instrument used in this study was the mathematics subtest of a multi-purpose placement test administered to all new undergraduate students as part of a freshman testing, counseling, and advising program.

The mathematics subtest is composed of 72 multiple choice questions that assess basic mathematical skills such as manipulation of numbers, reading graphs and tables, performing simple calculations, and using algebra, analytic geometry, trigonometry and calculus. Student answers are recorded on a special form that is scanned. Performance level on the test is used to guide students in their initial selection of mathematics courses.

Administration

Data for the math subtest was used in the parameterization. The exam was administered to groups of admitted students over the course of several months using uniform procedures. Time allowed for the test gave approximately 70 seconds per question (Examiner's Manual, 1992) which was sufficient for all students to attempt all questions (Leonard, 1992). Missing responses are therefore regarded as wrong instead of "not reached" for purposes of item calibration . A total of 14,914 students took the exam in 1991, however only 1921 students answered each item.

Data

Since one of the objectives of this study was to investigate the performance of the systems with and without omitted responses, two data matrices were formed. A sample of 1000 students was randomly drawn from each of two populations: the entire 14,914 cases and the 1,921 cases that had no responses omitted. This provided two data matrices of size 1000 cases by 72 responses, termed the samples with "complete responses" and "omitted responses."

Item Parameterization

The data were properly formatted for each program and run using the following program and computer combinations:

A. BILOG 3.04 (Mislevy & Bock, 1990) was run on an IBM Model 70 386 PC.

B. RASCAL 3.5 (1992) and ASCAL (MicroCAT, 1989) were run on a Unisys PW80 386 PC.

RASCAL used abilities scaled to a mean of 0 and a standard deviation of 1 so that those estimates are consistent with 2- and 3-parameter estimates. Expected *a posteriori* (EAP) estimates of ability were used for both MicroCAT and BILOG calculations because of the attractive statistical properties of this estimate (Bock and Mislevy, 1982).

Results

The analysis of results is organized into three parts: comparison of item parameter estimates, comparison of ability estimates, and comparison of item fit statistics. For each part, the results for each combination of model and sample are examined using SAS procedures (SAS, 1991).

Comparison of item parameter estimates

Table 1 contains descriptive statistics for item parameters as well as the correlation between the MicroCAT and BILOG estimates for each sample (complete responses and omitted responses) for the Rasch and 2 parameter models. Table 2 contains similar information for the 3 parameter model.

Insert Table 1 about here

Descriptive statistics for MicroCAT and BILOG estimates are nearly identical for the Rasch model.

For the two parameter model there is close agreement between the two programs for the sample with omitted responses, but less so for the sample with complete responses. In this latter sample, for example, the two estimates of discrimination correlated .93 while those for difficulty correlated .91.

For the three parameter model data contained in Table 2, the two estimates of discrimination and guessing each correlate about .90, while the difficulty parameter estimates correlate about .94

for the sample with complete responses indicating some disagreement between MicroCAT and BILOG estimates. For the sample with omitted responses, these correlations are somewhat higher.

Insert Table 2 about here

Of note also is the high degree of skewness in the distribution of BILOG estimates relative to MicroCat estimates under certain conditions. BILOG estimates of item difficulty in the 2- and 3-parameter model for the all responses completed sample and of the guessing parameter for both samples (applies only to the 3-parameter model) showed a much more highly skewed distribution. In each case the direction of the skewness for difficulty is negative while it is positive for guessing. A close examination of the distributions revealed that just three extreme values led to the skewness. MicroCAT uses both an upper and lower bound on these estimates and so eliminates the effect of outliers while BILOG does not. The large values observed for skewness for the BILOG estimates are a direct result of this difference.

Comparison of ability estimates

Table 3 displays descriptive statistics for each sample and model for the ability estimates.

Insert Table 3 about here

In addition, the correlation coefficients between the two estimates and between the raw score and each of the two estimates are also tabled. The mean and standard deviations are constrained to 0.00 and 1.00 respectively, so it is not surprising to observe values in these indices that are very similar for the two systems. There is remarkably high agreement between the BILOG and MicroCAT estimates as evidenced in the correlation coefficients equal to or very near 1.00 between the two estimates of ability and between the estimates and the raw score of the test takers. The correlations are quite high for the 1, 2 and 3 parameter models and for both samples alike. Careful examination of ability estimates plotted against raw score, however reveals regions of the

distribution where agreement between BILOG and MicroCAT is lower than one would be led to believe if only correlations were examined.

The relationship of the BILOG and the MicroCAT estimates with raw score is plotted in Figures 1 through 6. In each of these figures two plots are given side by side. The first plots the BILOG estimate against raw score while the right-hand plot in each figure plots the MicroCAT estimates against raw score. Figure 1 contains the plot of ability against raw score for the estimates based on a Rasch model and for the sample with omitted responses. Figure 2 plots similar data for the sample with complete responses. In both of these figures, it is evident that the relationship between ability estimates and raw score is very nearly linear and close to identical for MicroCAT and BILOG.

Insert Figures 1 and 2 about here

Figures 3 and 4 contain the plots for the 2 parameter model solution for omitted response and complete response samples respectively. Comparing the two figures, similar trends are evident. For each sample, the BILOG estimates appear to have a more nearly linear relationship with raw score than do MicroCAT estimates although this difference is not large. The shapes of the bivariate plots for BILOG and MicroCAT estimates are very similar when they are compared across samples indicating that the presence of omitted responses apparently did not affect the nature of the relationship of ability estimates with raw score.

Insert Figures 3 and 4 about here

Figures 5 and 6 display the bivariate plot of ability estimate from a 3-parameter model against raw score for the omitted response and the complete response samples, respectively. In these figures the BILOG ability estimates exhibit a more nearly linear relationship with raw score than do the MicroCAT estimates though the difference is not large. As with the 2 parameter model,

the presence of omitted responses did not change the relationship of ability to raw score for either program.

Insert Figures 5 and 6 about here

Comparison of Fit Statistics

Each procedure reported the value of a chi-square test statistic and its degrees of freedom for each test question. This statistic reflected the goodness of fit of the model to actual response data for each test question. In each program, the sample was divided into categories of equal size based on estimated ability. The proportion of examinees in each category who correctly and incorrectly answered each question are tabled. Expected values were calculated from the marginal values of this matrix. Large values of the chi-square statistic indicates a poor fit. Users of these programs examine these data to help identify items for deletion in a procedure aimed at improving the overall fit of the model to the response data.

Table 4 shows the numbers of items identified as "ill-fitting" for each program using a value of the chi-square test statistic that is significant at a p of .01. Applying this standard to both MicroCAT and BILOG gave an idea about how the two systems compare in identifying items for possible deletion from the test. The pattern was well-defined.

Insert Table 4 about here

For both the omitted response and the complete response samples, both programs agreed on the suspect items in approximately two-thirds of the items for the Rasch model. For the other models, one-third to one-half of the items were identified as "ill-fitting" by both programs. In general, BILOG seemed to give an impression of fitting the data better than MicroCAT does, since it identified a smaller number of suspect or weak items. The two programs differed to some extent in how ability categories were determined. These differences may have contributed to the

differences seen in chi-square test statistics. These differences between the two programs were relatively small for Rasch models, but more pronounced for 2- and 3-parameter models.

Conclusions and Implications

It was not surprising to discover imprecision at low ability levels since this has been a rather consistent finding in the literature. It was not surprising, either, to discover a relatively high level of agreement between BILOG and MicroCAT in estimating both item parameters and ability since the number of items is not especially small and the samples were relatively large. Differences between the two methods of making estimates are expected to produce differences when the number of items is small and the sample of examinees is small. In these situations, it is expected that BILOG would product better estimates (Mislevy and Stocking, 1989). What was surprising were the differences found in the estimates of ability of low scorers and the different decisions that would be taken in deleting items using a common cutoff value.

Based on the data and this research, the two programs appear to perform equally well for Rasch calibration. When 2-parameter calibration was chosen, BILOG appeared to show a more linear relationship with raw score, while the opposite was true for 3-parameter calibration. These differences were small with both programs showing high linear correlations. Choice of sample, all response or omitted response, showed no effect on the result.

When compared using the goodness of fit criteria, the BILOG program appeared to fit the data better than the MicroCAT program. For or the Rasch calibration, both programs agreed that most of the items were ill-fitting; thus for these data, Rasch calibration would not be suitable.

Recommendations for Future Research

Further research into the conditions surrounding the apparent reversal in the ability calculation for low ability students. Additionally, this research should be extended by deleting items as suggested by the goodness of fit criteria for each sample and model combination and then comparing the estimates of parameters and ability.

References

Bock, R. D. (1975). *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

Bock, R. D. & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443- 459.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrica, 35*, 179-197.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP Estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6(4)*, 431-444.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Examiner's Manual, Freshman Testing Program (1992). University Park, PA: The Pennsylvania State University.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory Principles and Applications*. Boston: Kluwer Nijhoff.

Hartwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13*, 243-271.

Jensma, C. J. (1976). A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement, 36*, 705-715.

Leonard, M. (1992). Personal Interview, 14 September 1992.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.

MicroCAT Testing System Users Manual (3rd. ed.) (1989). St. Paul: Assessment Systems Corporation.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3 Item Analysis and Test scoring with Binary Logistic Models*. Mooresville, In.: Scientific Software.

Mislevy, R. J. & Stocking, M. L. (1989). A consumers Guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

RASCAL Rasch analysis program user's Manual (1992). St Paul,MN: Assessment Systems Corporation.

SAS, Version 6 (1991). Cary, NC: The SAS Institute.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one- parameter logistic model. *Psychometrika*, 47, 175-186.

Tsutakawa, R. K. & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, 51, 251-263.

Vale, C. D. & Gialluca, K. A. (1985, November). *ASCAL: A microcomputer program for estimating logistic IRT parameters* (Research Report ONR-85-4). St. Paul, MN: Assessment Systems Corporation.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch Measurement*. Chicago: MESA Press.

Table 1

Comparison of item parameters

parameter	Mean	Std. dev.	Skew.	Pearson r
One parameter(Rasch) model				
Sample with omitted responses				
b(difficulty) MicroCAT	0.13	1.01	-0.68	1.00
b(difficulty) BILOG	0.16	1.08	-0.70	
Sample with all responses complete				
b(difficulty) MicroCAT	-0.29	1.05	-0.65	1.00
b(difficulty) BILOG	-0.28	1.15	-0.68	
Two parameter model				
Sample with omitted responses				
a(discrimination) MicroCAT	0.83	0.37	1.30	0.98
a(discrimination) BILOG	0.77	0.32	1.00	
b(difficulty) MicroCAT	0.21	1.05	-0.27	0.99
b(difficulty) BILOG	0.20	1.15	-0.64	
Sample with all responses complete				
a(discrimination) MicroCAT	0.84	0.36	0.94	0.93
a(discrimination) BILOG	0.79	0.33	0.71	
b(difficulty) MicroCAT	-0.21	1.02	-0.27	0.91
b(difficulty) BILOG	-0.26	1.38	-2.61	

Table 2

Comparison of item parameters

parameter	Mean	Std. dev.	Skew.	Pearson r
Three parameter model				
Sample with omitted responses				
a(discrimination) MicroCAT	1.25	0.41	0.29	0.97
a(discrimination) BILOG	1.23	0.43	0.27	
b(difficulty) MicroCAT	0.46	1.00	-0.84	0.99
b(difficulty) BILOG	0.45	1.02	-1.15	
c(guessing) MicroCAT	0.16	0.07	0.33	0.94
c(guessing) BILOG	0.13	0.07	1.20	
Sample with all responses complete				
a(discrimination) MicroCAT	1.22	0.45	0.65	0.90
a(discrimination) BILOG	1.77	0.40	0.17	
b(difficulty) MicroCAT	0.12	1.12	-0.37	0.94
b(difficulty) BILOG	0.06	1.26	-2.07	
c(guessing) MicroCAT	0.16	0.07	0.52	0.90
c(guessing) BILOG	0.17	0.08	1.35	

Table 3

Comparison of Ability Estimates (Theta)

	mean	Std. dev.	Skew.l	# correct	Correlation between meth.
One parameter(Rasch) model					
Sample with omitted responses					
EAP MicroCAT	0.00	0.93	0.63	1.00	1.00
EAP BILOG	0.00	1.00	0.62	1.00	
Sample with all responses complete					
EAP MicroCAT	-0.01	0.93	0.53	0.99	1.00
EAP BILOG	0.00	1.00	0.51	1.00	
Two parameter model					
Sample with omitted responses					
EAP MicroCAT	-0.02	0.98	0.13	0.96	0.97
EAP BILOG	0.00	1.00	0.74	0.99	
Sample with all responses complete					
EAP MicroCAT	-0.00	0.97	0.05	0.98	0.98
EAP BILOG	0.00	1.00	0.63	0.99	
Three parameter model					
Sample with omitted responses					
EAP MicroCAT	-0.01	0.95	0.80	0.99	0.97
EAP BILOG	0.00	1.00	-0.02	0.97	
Sample with all responses complete					
EAP MicroCAT	-0.01	0.95	0.67	0.99	0.97
EAP BILOG	0.00	1.00	-0.12	0.98	

TABLE 4
Number of Items Identified as Candidates for Deletion based on Criterion of Chi-Square $p < .01$

Data Set	All Responses			Omitted Responses		
Parameter Model	1	2	3	1	2	3
Program						
Both	34	6	0	32	11	2
BILOG Only	12	2	0	11	3	2
MicroCAT Only	5	10	7	4	7	2
Total	51	18	7	47	21	6

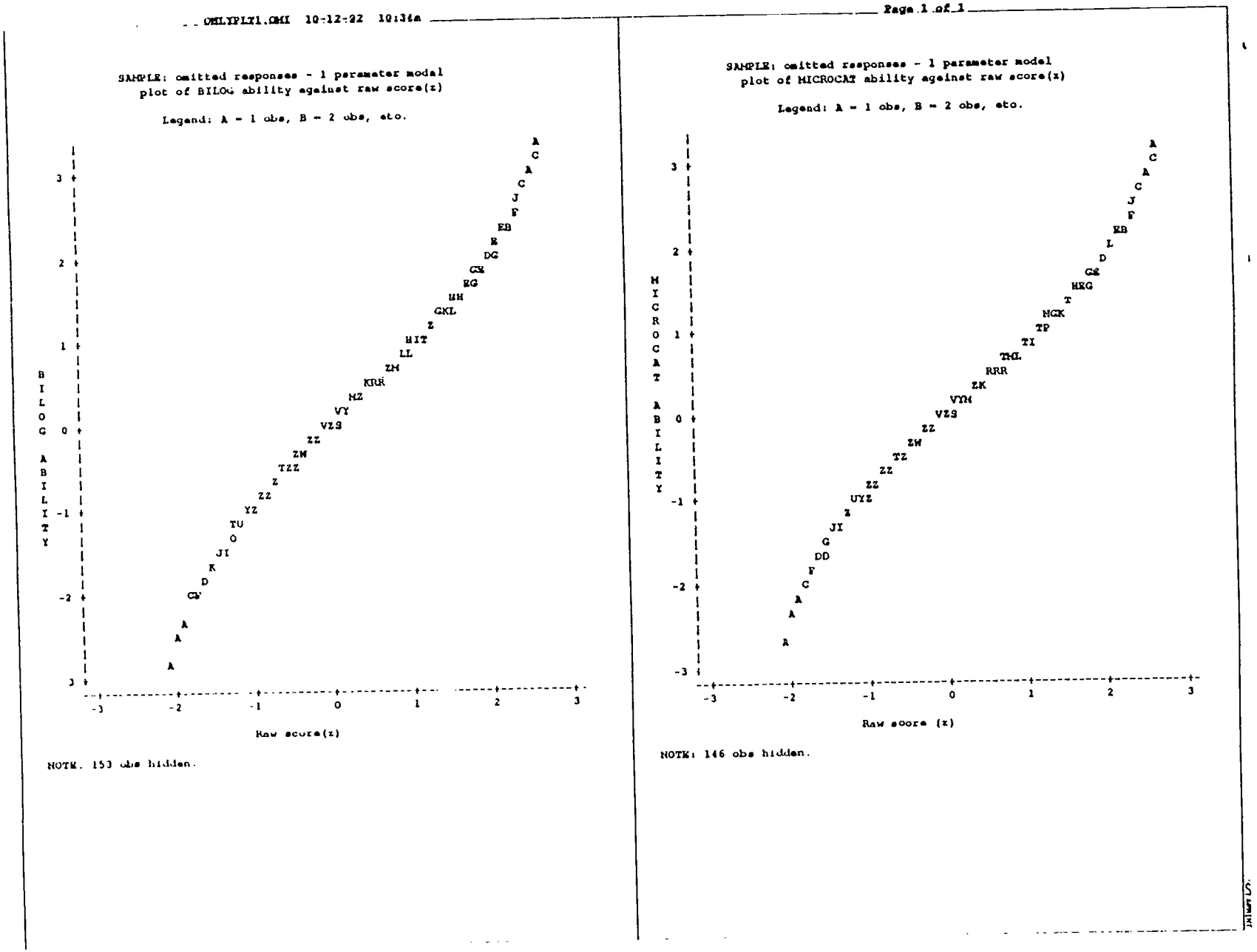

Figure 1 - Ability vs Raw Score, Rasch Model, Omitted Responses

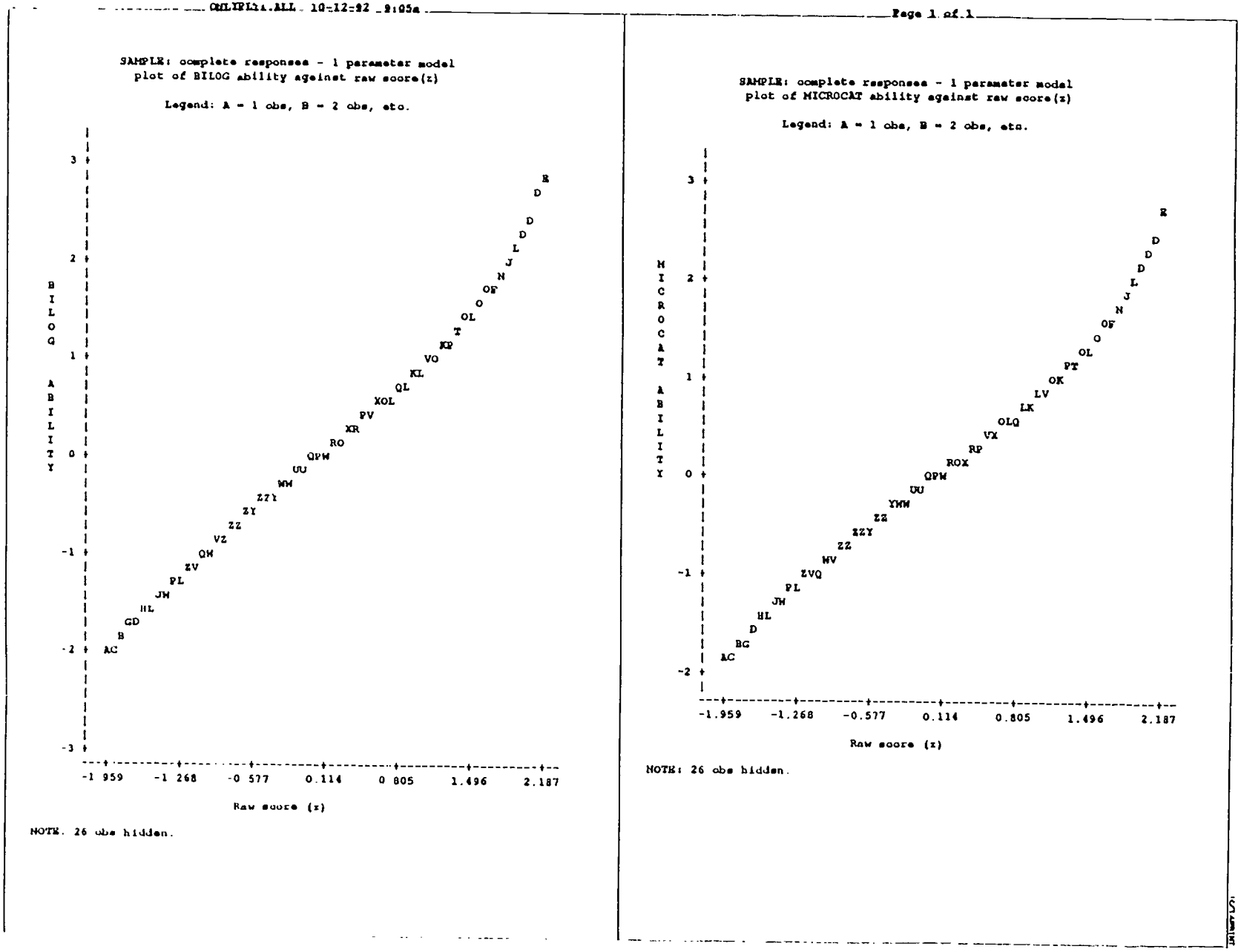

Figure 2 - Ability vs Raw Score, Rasch Model, All Responses

MicroCAT and BILOG 20

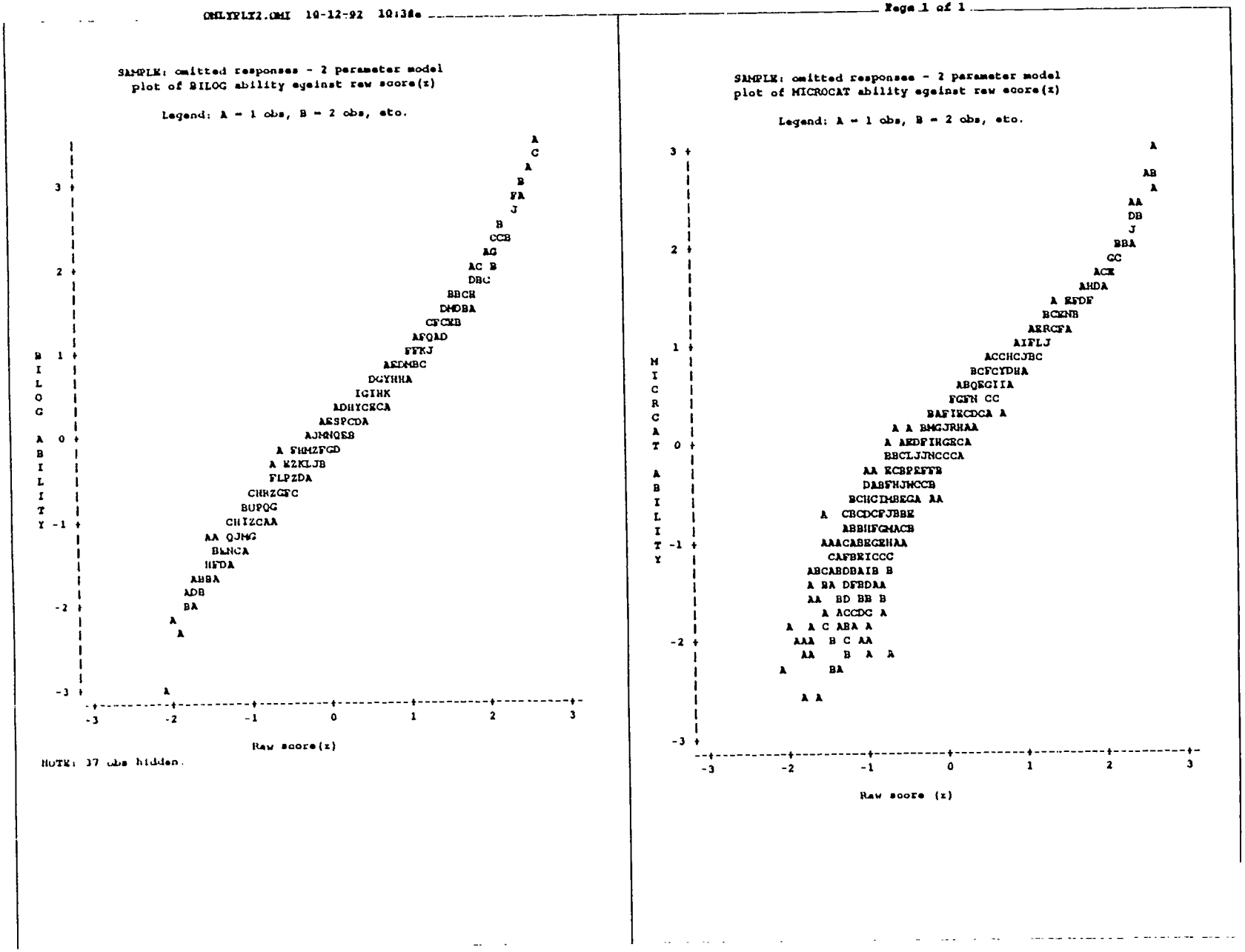

Figure 3 - Ability vs Raw Score, 2-Par Model, Omitted Responses

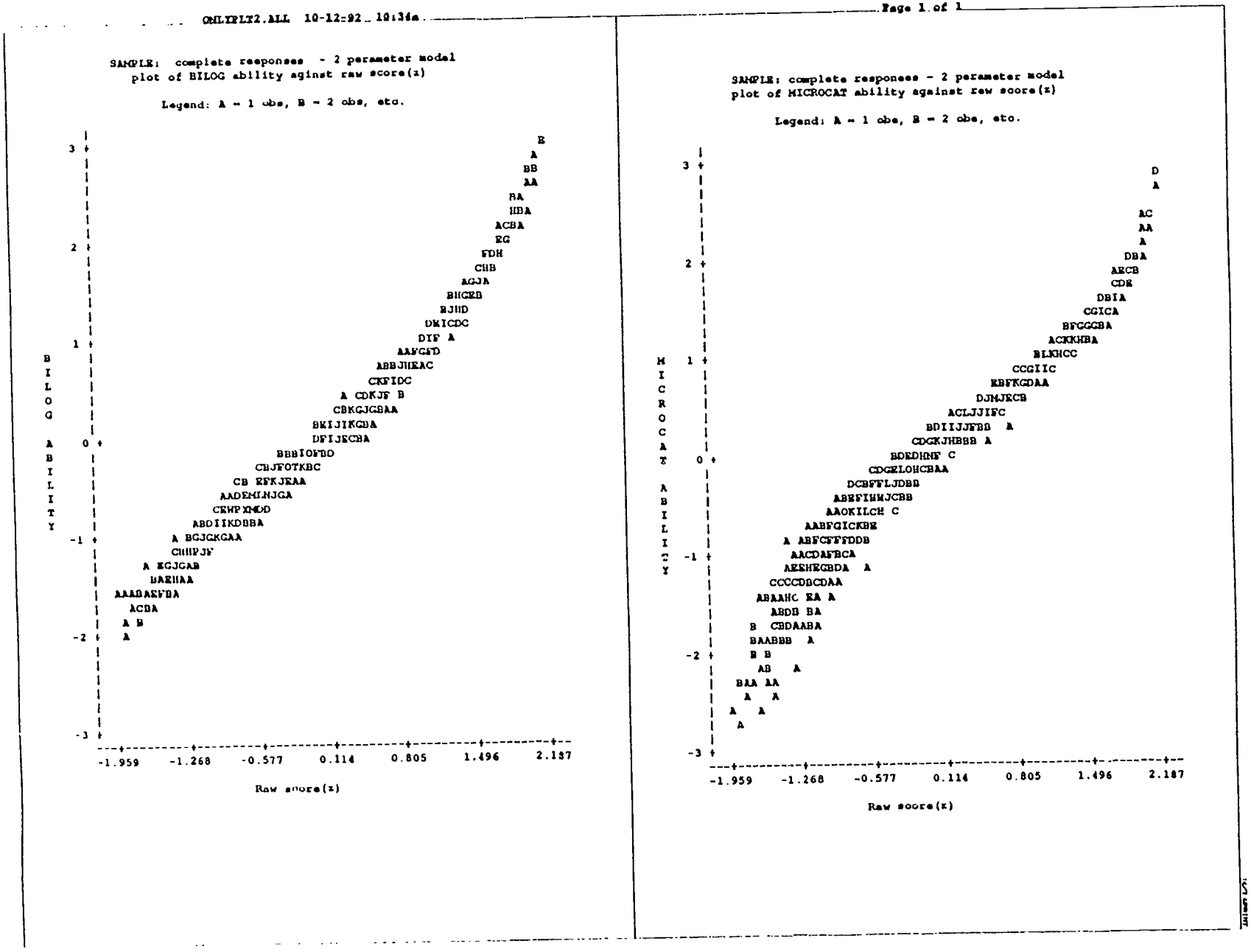

Figure 4 - Ability vs Raw Score, 2-Parameter Model, All Responses

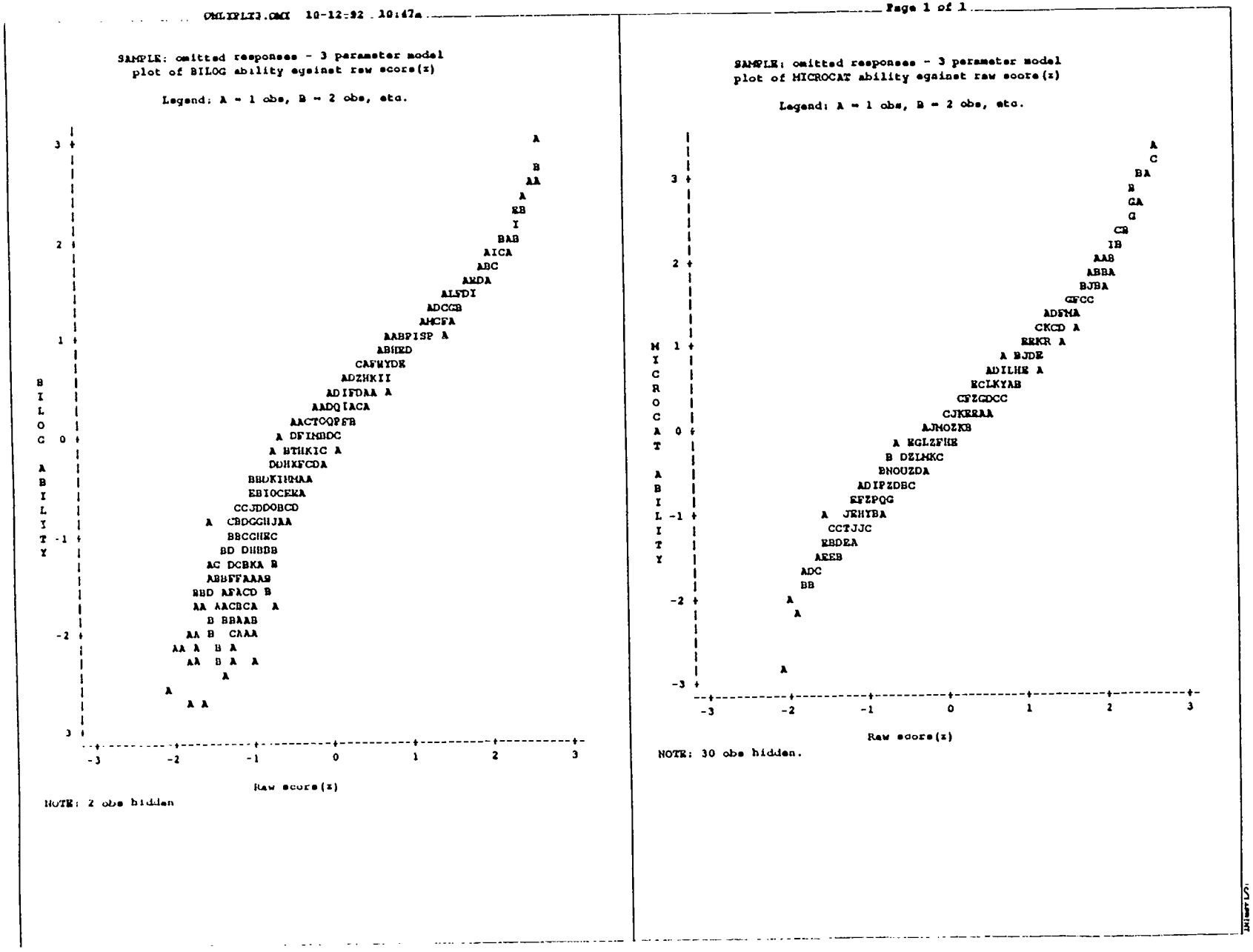

Figure 5 - Ability vs Raw Score, 3-Par Model, Omitted Responses

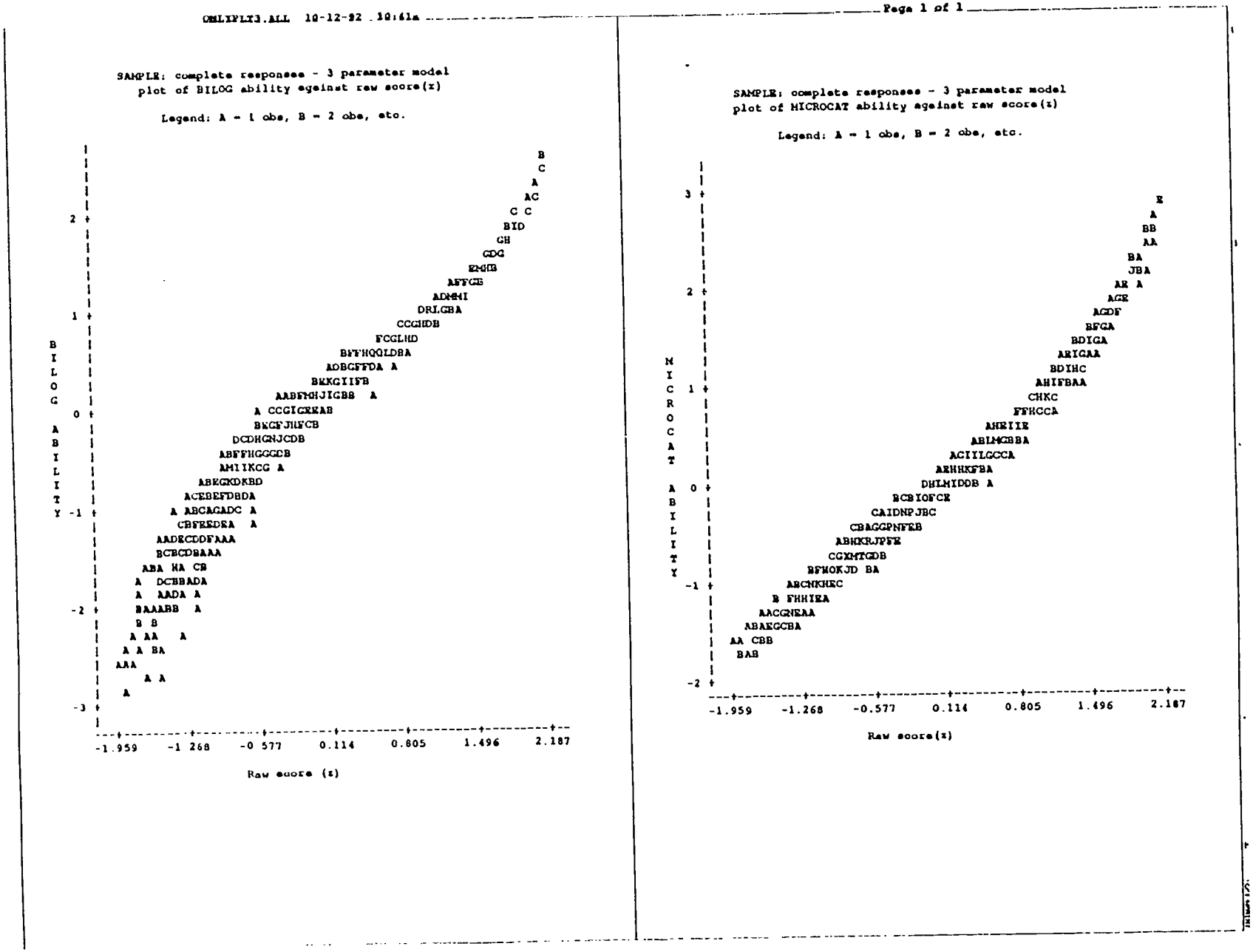

Figure 6 - Ability vs Raw Score, 3-Parameter Model, All Responses

MicroCAT and BILOG 24